Hypothesis_testing

November 27, 2024

1 Hypothesis testing principle

A hypothesis test assesses two mutually exclusive statements about any particular population and determines which statement is best established by the sample data. Here, we used two essential keywords: population and sample. A population includes all the elements from a set of data, whereas a sample consists of one or more observations taken from any particular population.

1.1 Hypothesis testing principle

Hypothesis testing is based on two fundamental principles of statistics, namely, normalization and standard normalization:

1. Normalization: The concept of normalization differs with respect to the context. To understand the concept of normalization easily, it is the process of adjusting values measured on different scales to common scales before performing descriptive statistics, and it is denoted by the following equation:

$$X_{changed} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

2. Standard normalization: Standard normalization is similar to normalization except it has a mean of 0 and a standard deviation of 1. Standard normalization is denoted by the following equation:

$$X_{changed} = \frac{X - \mu}{\sigma}$$

Besides these concepts, we need to know about some important parameters of hypothesis testing:

1. The null hypothesis is the most basic assumption made based on the knowledge about the domain. For example, the average typing speed of a person is 38-40 words per minute.
2. An alternative hypothesis is a different hypothesis that opposes the null hypothesis. The main task here is whether we accept or reject the alternative hypothesis based on the experimentation results. For example, the average typing speed of a person is always less than 38-40 words per minute. We can either accept or reject this hypothesis based on certain facts. For example, we can find a person who can type at a speed of 38 words per minute and it will disprove this hypothesis. Hence, we can reject this statement.
3. Type I error and Type II error: When we either accept or reject a hypothesis, there are two types of errors that we could make. They are referred to as Type I and Type II errors:

   A> False-positive: A Type I error is when we reject the null hypothesis (H0) when H0 is true.

   B> False-negative: A Type II error is when we do not reject the null hypothesis (H0) when H0 is false.
4. P-values: This is also referred to as the probability value or asymptotic significance. It is the probability for a particular statistical model given that the null hypothesis is true. Generally, if the P-value is lower than a predetermined threshold, we reject the null hypothesis.
5. Level of significance: This is one of the most important concepts that you should be familiar with before using the hypothesis. The level of significance is the degree of importance with which we are either accepting or rejecting the null hypothesis. We must note that 100% accuracy is not possible for accepting or rejecting. We generally select a level of significance based on our subject and domain. Generally, it is 0.05 or 5%. It means that our output should be 95% confident that it supports our null hypothesis.

1.2 Types of hypothesis testing

There are different types of hypothesis testing. The most commonly used ones are as follows: 1. Z-test 2. T-test 3. ANOVA test 4. Chi-squared test

Going through each type of test is beyond the scope of this book. We recommend checking out Wikipedia or the links in the Further reading section to get detailed information about them. However, we are going to look at the Z-test and the T-test in this book. In the preceding examples, we only used the Z-test.

1.2.1 T-test

The T-test is a type of test most commonly used in inferential statistics. This test is most commonly used in scenarios where we need to understand if there is a significant difference between the means of two groups. For example, say we have a dataset of students from certain classes. The dataset contains the height of each student. We are checking whether the average height is 175 cm or not:

1. Population: All students in that class
2. Parameter of interest: , the population of a classroom
3. Null hypothesis: The average height is $= 175$
4. Alternative hypothesis: > 175
5. Confidence level: $= 0.05$

In a study about mental health in Youth, some studies revealed 48% of parents believed that social media was the cause of their teenager's stress.

Population: Parent with a teenager (age >= 18)

Parameter of Interest: p

Null Hypothesis: p = 0.48

Aternative Hypothesis: p > 0.48

Data: 4500 people were surveyed and 65% of those who were surveyed believe that their teenagers' stress is caused due to social media.

1.2.2 p-hacking

p-hacking is a serious methodological issue. It is also referred to as data fishing, data butchery, or data dredging. It is the misuse of data analysis to detect patterns in data that can be statistically meaningful. This is done by conducting one or more tests and only publishing those that come back with higher-significance results.

1.3 Regression

We use correlation in statistical terms to denote the association between two quantitative variables.When it comes to quantitative variables and correlation, we also assume that the relationship is linear, that is, one variable increases or decreases by a fixed amount when there is an increase or decrease in another variable. To determine a similar relationship, there is the other method that's often used in these situations, regression, which includes determining the best straight line for the relationship. A simple equation, called the regression equation, can represent the relation:

$$Y = a + bX + u$$

Let's examine this formula:

Y = The dependent variable (the variable that you are trying to predict). It is often referred to as the outcome variable.

X = The independent variable (the variable that you are using to predict Y). It is often referred to as the predictor, or the covariate or feature.

a = The intercept.

b = The slope.

u = The regression residual.

If y represents the dependent variable and x represents the independent variable, this relationship is described as the regression of y on x. The relationship between x and y is generally represented by an equation. The equation shows how much y changes with respect to x. There are several reasons why people use regression analysis. The most obvious reasons are as follows: 1. We can

use regression analysis to predict future economic conditions, trends, or values. 2. We can use regression analysis to determine the relationship between two or more variables. 3. We can use regression analysis to understand how one variable changes when another also change.

1.3.1 Types of regression

The two main regression types are linear regression and multiple linear regression. Most simple data can be represented by linear regression. Some complex data follows multiple linear regression.

Simple linear regression Linear regression, which is also called simple linear regression, defines the relationship between two variables using a straight line. During linear regression, our aim is to draw a line closest to the data by finding the slope and intercept that define the line. The equation for simple linear regression is generally given as follows:

$$Y = a + bX + u$$

X is a single feature, Y is a target, and a and b are the intercept and slope respectively. The question is, how do we choose a and b? The answer is to choose the line that minimizes the error function, u. This error function is also known as loss or cost function, which is the sum of the square (to ignore the positive and negative cancelation) of the difference of the vertical distance between the line and the data point.

Multiple linear regression In the case of multiple linear regression, two more independent variables or explanatory variables show a linear relationship with the target or dependent variables. Most of the linearly describable phenomena in nature are captured by multiple linear regression. For example, the price of any item depends on the quantity being purchased, the time of the year, and the number of items available in the inventory. For instance, the price of a bottle of wine depends primarily on how many bottles you bought. Also, the price is a bit higher during festivals such as Christmas. Moreover, if there are a limited number of bottles in the inventory, the price is likely to go even higher. In this case, the price of wine is dependent on three variables: quantity, time of year, and stock quantity. This type of relationship can be captured using multiple linear regression. The equation for multiple linear regression is generally given as follows:

$$Y = a + b1X1 + b2X2 + b3X3 + \ldots + btXt + u$$

Here, Y is the dependent variable and Xis is the independent variable.

Nonlinear regression Nonlinear regression is a type of regression analysis in which data follows a model and is then represented as a function of mathematics. Simple linear regression relates to

two variables (X and Y) with a straight line function, , whereas nonlinear regression has to generate a curve. Nonlinear regression uses a regression equation, which is as follows:

$$Y = f(X, \beta) + \varepsilon$$

Let's look at this formula:

X = A vector of p predictors

= A vector of k parameters

f(-) = A known regression function

= An error term

Nonlinear regression can fit an enormous variety of curves. It uses logarithmic functions, trigonometric functions, exponential functions, and many other fitting methods. This modeling is similar to linear regression modeling because both attempt to graphically control a specific answer from a set of variables. These are more complicated to develop than linear models because the function is generated by means of a series of approximations (iterations) that may result from trial and error. Mathematicians use a variety of established methods, such as the Gauss-Newton and Levenberg-Marquardt methods. The goal of this nonlinear model generated curve line is to make the OLS as small as possible. The smaller the OLS the better the function fits in the dataset's points. It measures how many observations vary from the dataset average.

[ ]:
```
n = 4500
pnull= 0.48
phat = 0.65
```

[ ]:
```
import statsmodels.api as sm
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
```

[ ]:
```
sm.stats.proportions_ztest(phat * n, n, pnull, alternative='larger')
```

[ ]:
```
(23.90916877786327, 1.2294951052777303e-126)
```

Our calculated p-value is 1.2294951052777303e-126 is pretty small and we can reject the Null Hypothesis (H0).

[ ]:
```
import numpy as np

sdata = np.random.randint(200, 250, 89)
sm.stats.ztest(sdata, value = 80, alternative = "larger")
```

[ ]: (96.71588016677123, 0.0)

[ ]:
```
sm.stats.ztest(sdata, value = 80, alternative = "larger")
```

[ ]: (96.71588016677123, 0.0)

2 T-test

[ ]:
```
height = np.array([172, 184, 174, 168, 174, 183, 173, 173, 184, 179, 171, 173,␣
 ↪181, 183, 172, 178, 170, 182, 181, 172, 175, 170, 168, 178, 170, 181, 180,␣
 ↪173, 183, 180, 177, 181, 171, 173, 171, 182, 180, 170, 172, 175, 178, 174,␣
 ↪184, 177, 181, 180, 178, 179, 175, 170, 182, 176, 183, 179, 177])
height
```

[ ]:
```
array([172, 184, 174, 168, 174, 183, 173, 173, 184, 179, 171, 173, 181,
       183, 172, 178, 170, 182, 181, 172, 175, 170, 168, 178, 170, 181,
       180, 173, 183, 180, 177, 181, 171, 173, 171, 182, 180, 170, 172,
       175, 178, 174, 184, 177, 181, 180, 178, 179, 175, 170, 182, 176,
       183, 179, 177])
```

[ ]:
```
from scipy.stats import ttest_1samp
import numpy as np


height_average = np.mean(height)
print("Average height is = {0:.3f}".format(height_average))

tset,pval = ttest_1samp(height, 175)

print("P-value = {}".format(pval))

if pval < 0.05:
  print("We are rejecting the null Hypotheis.")
else:
  print("We are accepting the null hypothesis")
```

```
Average height is = 175.618
P-value = 0.35408130524750125
We are accepting the null hypothesis
```